IT550: Information Retrieval

Assignment 5: Document Representation (Part 3)

Instructor: Prasenjit Majumder

Learning Outcome: At the end of this assignment you will learn about Word2Vec which is used representing words in embedding space. And representation of document using word embedding.

1 Problem description

Word embedding is representation of words in latent space such that similar words are grouped together. In our case we will deal with distributed representation where we will learn the vector representation of a word based on words surrounding it. One method that uses distributed representation for words is Word2Vec introduced by Mikolov et al. Word2Vec has two variations: Continuous Bag of Words (CBOW) and Skip Gram. In this assignment we will use CBOW and Skipgram for obtaining the distributed representation.

2 Implementation

2.1 Dataset

- For this assignment we will use Telegraph news articles, which is in XML format. It contains news on different categories for the year 2004 to 2007. You can download the dataset from this link: https://drive.google.com/open?id=1JuawXQmYVkjpfL3HOblqjDrqw8V1lHrC
- The Queries are in "en.topics.76-125.2010". The query is of the format shown in Figure 1. Use the sentences enclosed in desc tag for framing your query vector
- The "en.qrels.76-125.2010.txt" contains the documents that are relevant to a query. The format of a qrel is such: Query_No Q0 Document ID Relevance score.
- Relevance score is binary 0 or 1. 1 is for relevant, 0 is for otherwise.
- The documents in the dataset is in the format shown in Figure 2.

```
<top lang='en'>
<num>76</num>
<title>Clashes between the Gurjars and Meenas</title>
<desc>
Reasons behind the protests by Meena leaders against the
inclusion of Gurjars in the Scheduled Tribes.
</desc>
<narr>
The Gurjars are agitating in order to attain the status of a
Scheduled Tribe. Leaders belonging to the Meena sect have
been vigorously opposing this move. What are the main reasons
behind the Meenas' opposition? A relevant document should
mention the root cause(s) behind the conflict between these
two sects.
</narr>
</top>
```

Figure 1: Query Format

2.2 Exercise

1. Implement CBOW and Skipgram approach using Gensim library.
2. Represent document using word2vec representation.
3. Get vectors using Skip gram and CBOW
4. Use Mean Average Precision (MAP) for calculating your system performance for both CBOW and Skipgram approach.

```
<DOC>
<DOCNO> </DOCNO>
<TEXT></TEXT>
</DOC>
```

Figure 2: News Format

3 References

- Tomas Mikolov et al.,Distributed Representations of Words and Phrases and Their Compositionality, NIPS, Volume 2, 2013, 31113119
- https://pytorch.org/tutorials/beginner/nlp/word_embeddings_tutorial.html
- https://kavita-ganesan.com/gensim-word2vec-tutorial-starter-code/#.XlOEPigzZPY
- https://radimrehurek.com/gensim/models/word2vec.html
- http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

4 Submission

- You have to submit your assignment in Jupyter notebook with proper comments and explanation of your approach.
- In the end you will have to specify how many queries fetched the relevant document.
- The submission deadline for this assignment in **9th Mar 2021 at 11 PM**